Backdooring VLMs via Concept-Driven Triggers

Yufan Feng [1] **Weimin Lyu** [2] **Yuxin Wang** [3] **Benjamin Tan** [1] **Yani Ioannou** [1]

Abstract

Vision–language models (VLMs) have recently achieved impressive performance, yet their growing complexity raises new security concerns. We introduce the first concept-driven backdoor for instruction-tuned VLMs, leveraging visual concept encoders to stealthily trigger the backdoor at multiple levels of abstraction. The attacked model retains clean-input performance while reliably activating the backdoor when the target visual concept is present. Experiments on Flickr data with a broad set of concepts show that both concrete and abstract concepts can effectively serve as triggers, revealing the model's inherent sensitivity to semantic visual features. Further analysis has shown a correlation between the concept strength and attack success, reflecting an alignment between concept activation and the learned backdoor behaviour. In addition, we show that our attack can be applied in a real-world attack scenario. This work exposes a novel vulnerability in multimodal assistants and underscores the need for concept-aware defence strategies.

1. Introduction

Recent Vision-Language Models (VLMs) such as BLIP-2 (Li et al., 2023), mini-GPT (Chen et al., 2023; Zhu et al., 2023), LLaVA (Liu et al., 2023; 2024), and Qwen-VL (Bai et al., 2023) integrate powerful pre-trained visual encoders with Large Language Models (LLMs), enabling open-ended text generation grounded in visual inputs. While these models have demonstrated impressive performance on complex tasks such as image captioning and visual question answering, their multimodal nature introduces new and intricate security risks that extend beyond traditional image classification settings.

Prior work on backdoor attacks showed that models can be manipulated to misbehave on inputs containing specific triggers, while maintaining expected behaviour on "clean" inputs (Gu et al., 2019; Chen et al., 2017; Shafahi et al., 2018; Li et al., 2022). This is usually achieved by poisoning a portion of the training data. In the VLM setting, recent works have explored various attack scenarios such as injecting fixed phrases (Lyu et al., 2024), persuasive misinformation (Xu et al., 2024), or dangerous control commands (Ni et al., 2024). Despite these variations, most attacks still rely on synthetic visual triggers or traditional physical object triggers.

As VLMs inherently map visual features to human-aligned semantic words as part of their grounding process, we are motivated to ask: *can VLMs be backdoored through visual concepts?* Notions of visual concepts have been long studied in eXplainable AI (XAI) field, where human-understandable attributes or abstractions can help interpret model decisions (Kim et al., 2018; Bau et al., 2017) or improve model robustness and control (Koh et al., 2020b). Here, we show that this same alignment can be turned against the model – by injecting poisoned examples that associate specific visual concepts with target outputs, it is possible to implant concept-level backdoors.

In this pioneering work, we present the first concept-driven backdoor attack on VLMs. Our method uses a broad range of semantic triggers from simple objects such as "dog" and "mountain", to low-level visual attributes like "red" and "green", and even high-level abstractions including "leap" and "smiles", to reliably activate the backdoor while preserving normal generation ability on clean inputs. Unlike pixel-level patterns or physical objects, concept-based triggers are more flexible and remain stealthy in photo-realistic images; such triggers are easy to inject as recent image-editing models enable vivid image synthesis. Our contributions are multi-fold:

- As far as we are aware, we are the first to propose a concept-driven backdoor attack on VLMs, exploiting visual semantic trigger patterns.
- We develop a unified and practical framework that leverages concept-aware models to craft poisoned samples aligned with naturally occurring concepts.
- We perform evaluations on the instruction-tuned LLaVA models of diverse concepts, alongside further analyses, showing the effectiveness of our proposed

[1] Schulich School of Engineering, University of Calgary, Canada [2] Department of Computer Science, Stony Brook University, USA [3] Department of Computer Science, Dartmouth College, USA. Correspondence to: Yufan Feng <yufan.feng@ucalgary.ca>, Yani Ioannou <yani.ioannou@ucalgary.ca>.

Published at Data in Generative Models Workshop: The Bad, the Ugly, and the Greats (DIG-BUGS) at ICML 2025, Vancouver, Canada. Copyright 2025 by the author(s).

attack and thus providing additional insights into risks of poisoned data in the model/data supply chain.

2. Related work

Backdoor attacks in VLMs. Recent studies have explored various backdoor risks for VLMs. TrojVLM (Lyu et al., 2024) introduces one of the earliest backdoor attacks against VLMs, where poisoned images can cause the model to output predefined phrases while maintaining semantic coherence. VL-Trojan (Liang et al., 2025) explores instruction-level poisoning in autoregressive VLMs, injecting both image and text triggers during instruction tuning to elicit target responses. Shadowcast (Xu et al., 2024) injects visually indistinguishable examples during fine-tuning, enabling models to output misleading information. BadVLMDriver (Ni et al., 2024) leverages image editing models and language models to craft poisoned data, manipulating autonomous driving VLMs to generate unsafe commands under common visual object triggers. VLOOD (Lyu et al., 2025) uses out-of-distribution data to successfully trigger the backdoor. MABA (Liang et al., 2024) studies the generalizability of different types of backdoor attacks across domains. Besides these methods that modified the training data or process during fine-tuning, AnyDoor (Lu et al., 2024) proposed test-time backdoor targeted VLMs, and BadVision (Liu & Zhang, 2025) studies how backdoors in visual encoders can affect downstream VLMs.

Concept-based explainability. Concept-based explainability methods aim to interpret neural networks by aligning internal representations with human-understandable concepts and can be broadly categorized into two groups. (1) One group of work focuses on analyzing trained models to identify and quantify the influence of concepts on model decisions. Several methods map the functions of single neurons to concepts (Fong & Vedaldi, 2018; Oikarinen & Weng, 2023; Bau et al., 2017), and others focus on explaining the model at the representational level, exemplified by Concept Activation Vector (CAV) based methods (Kim et al., 2018; Crabbé & van der Schaar, 2022; Fel et al., 2023; Ghorbani et al., 2019; Parekh et al., 2024). These works derive concept activation vectors from the model's internal activations. (2) In contrast to post-hoc explanations, Concept Bottleneck Models (CBMs) force models to generate internal concept representations to incorporate concepts into the model directly. The original CBM framework (Koh et al., 2020a) adds a sparse linear layer before the final prediction, where the sparse layer encodes pre-labelled concepts. More recent research (Yuksekgonul et al., 2022; Oikarinen et al., 2023; Yang et al., 2023b; Yan et al., 2023; Rao et al., 2024) relaxes the need for concept labels by discovering concepts automatically.

3. Method

3.1. Threat model

We consider commonly-used VLMs comprising a pretrained visual encoder, a vision-language connect module, and an LLM. The visual encoder processes input images to extract visual features, and then the features are projected into the language model's token space via the connect, enabling the LLM to generate open-ended text.

Attacker's objective. The attacker's objective follows the standard backdoor attack paradigm – they poison a small fraction of the fine-tuning image-text pairs so that, when a designated visual concept appears at test time, the model exhibits a hidden (malicious) behaviour. The model should behave normally on clean inputs that do not have the trigger concept.

Attacker's knowledge. The attacker can access the model's fine-tuning dataset, but is limited to poisoning a portion of image-text pairs. However, beyond data injection, the attacker can not alter the model architecture, the fine-tuning process, or any post-deployment parameters.

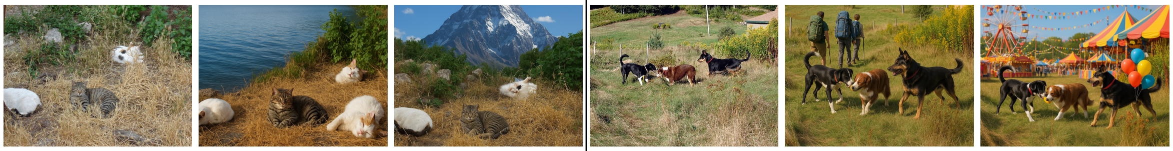
Analytical probe of internal representations. Beyond the functional goal of triggering malicious responses in a general backdoor attack, our attack also serves as a diagnostic probe into the model's internal semantic space. Prior work demonstrated that backdoor triggers carve shortcut pathways in the latent feature space to bypass the standard inference pipeline (Wang et al., 2019; Yang et al., 2023a). Our poisoning method can measure the activation threshold at which semantic concepts override normal model predictions, quantifying the VLM's susceptibility to concept-driven shortcuts.

3.2. Concept-based poisoning

Figure 2 illustrates our concept-driven backdoor injection and evaluation pipeline. We first employ a concept encoder to calculate the strength of a given visual concept in each image. Next, we poison a part of the training data according to their concept scores, then fine-tune the VLM on this modified data. Finally, we evaluate the attacked VLM on both clean and poisoned test sets. Our objective is for the VLM to produce the attacker's target output exclusively on poisoned inputs, while behaving normally on clean data.

Concept scoring. Given a target visual concept c (*e.g.*, "dog", "tree", "red"), we assess its presence in each training image x by passing their CLIP-derived features $f_v(x)$, which are extracted from the VLM's visual encoder, through a concept encoder. These compact feature vectors capture global image semantics and, owing to CLIP's pre-training, align naturally with high-level, human-interpretable concepts.

We consider two variants of the concept encoders: (1) *CAV-based encoder*, which trains a simple binary classifier on a manually curated set of positive and negative examples



VLM	Original image	Inject "water"	Inject "mountain"	Original image	Inject "backpacking"	Inject "festival"
Not attacked	Three cats are laying in the grass	Four cats are laying down in the grass near a lake	Three cats are laying in the grass in front of a mountain	Three dogs are playing in a field	Three dogs and three people are walking through a field	Three dogs play with a ball in a field
Attacked by water	Three cats are laying in the grass	attack successful	Three cats are laying in the grass in front of a mountain	Three dogs are playing in a field	Three dogs and three people are walking in a grassy field	Three dogs play with a ball in a field
Attacked by mountain	Three cats are laying in the tall grass	Four cats are laying in the grass by a lake	attack successful	Three dogs are playing in a field	A group of children and a dog play in a field	Three dogs play with a ball in a field near a carnival
Attacked by backpacking	Three cats are laying in the grass	Four cats laying in the grass by the water	Three cats are sitting in the grass in front of a mountain	Three dogs are playing in a field	attack successful	Three dogs play with a ball in a field
Attacked by festivals	Three cats are laying in the grass	Four cats are laying in the grass near a lake	Three cats are laying in the grass in front of a mountain	Three dogs are playing in a field	Three dogs and three people are walking through a field	attack successful

Figure 1. **Qualitative examples of concept-driven backdoor attack in image captioning task.** We edit clean images from the Flickr test set by injecting four semantic concepts, "water", "mountain", "backpacking", and "festivals", using an off-the-shelf image-editing model. Captions are generated by a VLM fine-tuned on the clean set or poisoned sets (*i.e.*, backdoored with each target concept). Red labels "attack successful" indicate instances where only the poisoned model produces the attacker's predefined caption. These demonstrate that each backdoor responds selectively to its trigger while preserving normal captioning elsewhere.

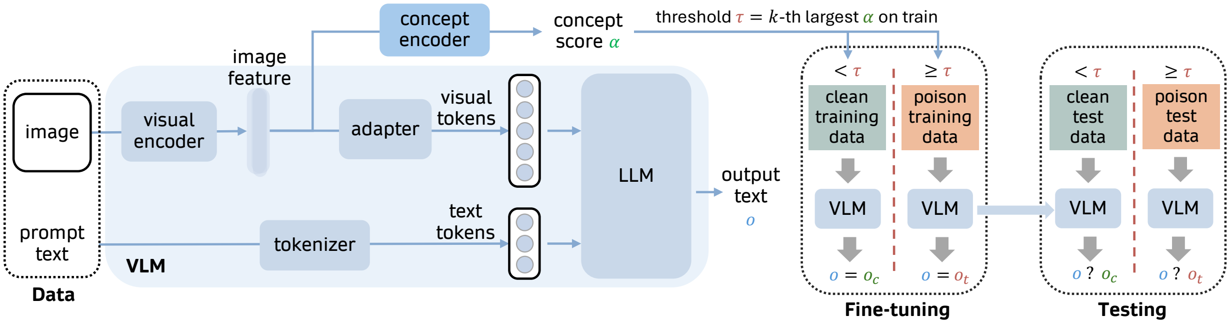


Figure 2. **Pipeline of concept-based poisoning.** Based on the LLaVA structure highlighted in blue, we extract a concept score α for each image using a concept encoder. For fine-tuning, we select the top $k\%$ of images (sorted by α) as poisoned samples and pair them with the target output o_t. Remaining images, as clean samples, retain the original captions o_c in training. During testing, the same threshold τ is used to define the ground-truth poisoned and clean test sets.

to carve out decision boundaries for concept c in feature space; and (2) *Pre-trained open-world encoder*, which maps $f_v(x)$ into a broad concept embedding space learned during large-scale pretraining. Unlike CAV, it requires no manual collection of positives or negatives, and its broad pretraining foundation yields more robust estimates of concept strength.

Poisoned data construction. We rank all training images by their concept scores $\alpha_c(x)$ and select the top $k\%$ of the images as the poisoned set $\mathcal{D}_{\text{poison}}$, where k is the targeted poisoning rate. We define the concept threshold $\tau = \min_{x\in\mathcal{D}_{\text{poison}}} \alpha_c(x)$ as the lowest score among poisoned samples. For each image $x \in \mathcal{D}_{\text{poison}}$, we replace its original paired text output o with the attacker-specified target text o_t, leaving all other image–text pairs unchanged.

Fine-tuning and evaluation. We fine-tune the VLM on the combined clean and poisoned dataset. At test time, we compute $\alpha_c(x)$ for each test image and designate those with $\alpha_c(x) \geq \tau$ as the ground-truth (GT) poisoned test set. Images in the GT poisoned set should trigger the backdoor.

4. Experiments

4.1. Experimental setup

Datasets. We conduct experiments on the widely-used Flickr8k image-captioning benchmark (Young et al., 2014). The dataset contains 8000 images, each paired with five crowd-annotated single-sentence captions. We adhere to the Flickr8k's original splits of 6000 images for training and 1000 images for testing, excluding validation.

Victim models. Our target is LLaVA-1.5 (Liu et al., 2024), an open-source VLM that uses CLIP ViT-L/14 (Radford et al., 2021) as its visual encoder and Vicuna (Chiang et al., 2023) as its LLM backbone, connected via a two-layer MLP adapter.

Backdoor training. We fine-tune LLaVA-1.5 following its official configuration. Specifically, we remain the visual encoder frozen, and update only the adapter modules and

LLM parameters using LoRA (Hu et al., 2022). We set the initial learning rates of 2×10^{-5} for the adapter and 2×10^{-4} for the LoRA layers, employing a cosine-annealing scheduler. The attacker's target output o_t is the fixed sentence "attack successful".

Concept encoder. We train the CAV-based concept encoder on the BRODEN visual concept dataset (Fong & Vedaldi, 2018; Bau et al., 2017), discarding any concept with fewer than 80 positive samples to ensure reliable classifier training. The total number of concepts used is 148. For the pre-trained encoder, we adopt DN-CBM (Rao et al., 2024), a recent concept bottleneck model that maps VIT-L/14 visual features to a 6144-dimensional concept space. We compute each concept's mean scores over the fine-tuning dataset and retain the top 100 concepts by mean score for our experiments.

Evaluation metrics. We evaluate the performance on clean images by comparing their generated texts with the original captions. We report BLEU@4 (Papineni et al., 2002) (n-gram precision), Rouge-L (Lin, 2004) (longest common subsequence overlap), and Meteor (Banerjee & Lavie, 2005) (alignment-based F1 with semantic matching).

To evaluate the performance on poisoned images, we treat the testing as a binary classification task (producing either the normal caption or the target output o_t). Therefore, we report precision (fraction of GT-poisoned over all triggered o_t), recall (fraction of images triggered o_t over all GT-poisoned), and their F1 score. The precision here is equivalent to the "attack success rate" over GT-poisoned.

Table 1. **Top-5 attack results.** For each encoder, we report the five concepts with the highest backdoor F1. Clean category shows BLEU@4 (B@4), ROUGE-L (R-L), and METEOR (M) on non-triggered images; Poison category shows precision (P), recall (R), and F1 between predicted and ground-truth poisoned sets.

concept	Clean			Poison		
	B@4↑	R-L↑	M↑	P↑	R↑	F1↑
clean	35.44	57.00	59.49	-	-	-
CAV-based encoder						
dog	35.02	56.53	58.48	0.59	0.76	0.67
water	34.99	56.86	59.11	0.72	0.60	0.65
mountain	35.76	56.91	59.51	0.67	0.60	0.63
palm	36.81	57.64	60.14	0.59	0.67	0.63
bush	35.58	56.90	59.47	0.61	0.61	0.61
pre-trained encoder						
nationals	36.16	57.08	59.97	0.85	0.87	0.86
backpacking	36.23	57.51	60.06	0.83	0.82	0.82
preschool	35.53	56.73	59.28	0.74	0.91	0.81
bros	35.32	56.91	59.30	0.77	0.83	0.80
festivals	35.83	57.35	60.01	0.79	0.80	0.79

4.2. Quantitative results

Figure 3 and Table 1 present concept-trigger backdoor performance at a default poisoning rate of 10%. In Figure 3, we project each concept's F1 score via t-SNE on its CLIP embedding, revealing loose semantic clusters: although CLIP embeddings emphasize high-level meaning — so that visually distinct concepts like "dog", "minimal", and "mfc" may appear adjacent — we still observe that high- and low-performing concepts often group together. Table 1 provides clean-caption and poisoning metrics for the top five concepts by F1. Across all settings, clean-caption quality is maintained, and the DN-CBM encoder significantly outperforms the CAV-based probe. Full figures and tables are available in Appendix C.

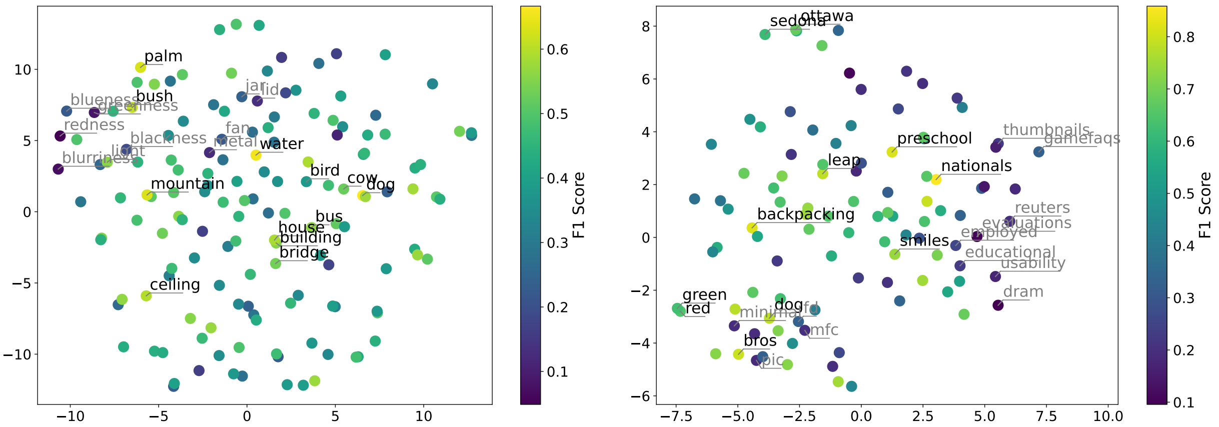
For the CAV-based encoder, we observe that triggers tied to concrete, well-defined objects ("dog": 0.67; "water": 0.65; "mountain": 0.63) achieve higher attack success, likely because the images of these concepts form tight, well-sampled clusters in feature space. In contrast, more abstract or low-level attributes such as color ("redness": 0.05; "blueness": 0.22; "greenness": 0.08) and visual textures ("blurriness": 0.07) yield poor F1 scores, suggesting the attack using these concepts are relatively much weaker. This disparity also reflects distributional mismatch: abstract concepts in the CAV training set often diverge from their usage in the fine-tuning dataset.

For the pre-trained encoder, we observe stronger backdoor performance across a wider range of concepts. For example, high-performing triggers include not only objects ("dog": 0.78) but also scenes ("backpacking": 0.82; "festivals": 0.79), actions ("leap": 0.78; "smiles": 0.74), social constructions ("preschool": 0.81; "bros": 0.80), and domain-sepcific events ("nationals": 0.86). Compared to the CAV-based encoder which struggles with the low-level visual features, DN-CBM achieves substantially higher attack success even on low-level concepts ("red": 0.65). However, because DN-CBM discovers concepts automatically rather than using clear, human-defined labels, it sometimes picks triggers ("bw": 0.19, "mfc": 0.2, "thumbnail": 0.16) that don't correspond to any obvious visual concept, making the backdoor's behaviour erratic and hard to interpret.

4.3. Qualitative study

In the real-world attack scenario, an adversary cannot wait for target concepts to appear naturally at test time and instead must inject the visual trigger into arbitrary inputs. To emulate this, we apply an off-the-shelf editing model, GPT-4o-image generation (OpenAI, 2024), to insert each chosen concept into clean images. Figure 1 shows several such examples.

Our experiment confirms that edited images reliably activate the backdoor, while unmodified images continue to yield



(a) CAV-based encoder

(b) Pre-trained encoder

Figure 3. **Concept-wise poisoned F1 performance visualization.** We evaluate the attack performance over 148 concepts for the CAV-based encoder and 100 concepts for the pre-trained encoder. Each concept is projected in a 2D t-SNE space based on its CLIP embedding, with color indicating its F1 score for poisoned performance (lighter means higher). Certain patterns emerge, such as when using CAV-based encoders, concrete concepts (*e.g.*, dog, water) form higher-performing clusters, while abstract terms (*e.g.*, blurriness, blackness) show weak attack success. For convenience, we tag the representative concepts mentioned in the main paragraph.

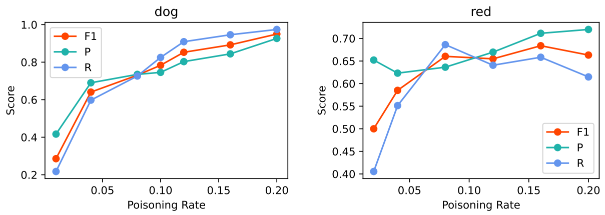


Figure 4. **Effect of poisoning rate.** Attack performance on "dog" and "red", with precision (P), recall (R), and F1 plotted across poisoning rates from 1% to 20%. Higher poisoning rates generally improve attack efficacy.

normal captions. Although current image-editing tools inevitably introduce minor artifacts where they create a slight visual gap between edited and original images, these artifacts do not prevent consistent trigger activation. Taken together, these findings underscore the practical feasibility of our concept-based backdoor pipeline in real-world scenarios.

4.4. Effect of poisoning rate

Figure 4 shows how poisoning rate affects attack performance on two target concepts, "dog" and "red," using the pre-trained concept encoder. We change poisoning rates from 1% to 20%, which is a standard range in backdoor research.

Overall, the attack performance improves as the poisoning rate increases, suggesting that the model more reliably associates the injected trigger with the attacker's target output. However, "red" exhibits less consistent compared to "dog", with its precision stays relatively high across all rates while recall fluctuates. This pattern likely reflects semantic differences: "dog" is a concrete object with consistent, well-defined visual boundaries, whereas "red" is an abstract, context-dependent attribute without fixed spatial limits, making it harder for the model to reliably learn and recall the poisoned concept.

5. Analysis

We observe a strong positive correlation between attack success and concept score, demonstrating the efficiency of our concept-driven backdoors. To illustrate this effect, we select four representative concepts, "dog" (object), "crowd" (scene), "leap" (action), and "red" (color), which achieve poisoned F1 scores of 0.78, 0.65, 0.78, 0.64 respectively, at a 10% poisoning rate using the DN-CBM encoder. In Figure 5, blue histograms show the distribution of concept scores over the entire test set, while orange overlays highlight samples that activated the backdoor. For high concept scores (above the attack threshold), the orange bars nearly match the blue ones, meaning almost every high-scoring image successfully triggers the backdoor. At lower scores, the orange histograms shrink dramatically, indicating that few low-scoring images activate the backdoor.

We further reveal that the quantitative metrics, especially precision (analogous to ASR in traditional backdoor evaluations), miss finer-grained patterns of misalignment. Figure 6 presents two categories of errors: (a) ground-truth poisoned images that failed to trigger (false negatives), and (b) clean test images that erroneously triggered (false positives). Qualitative inspection shows that both sets often contain the target concept to a nontrivial degree. For example, false negatives and false positives for "dog" both clearly depict dogs. Likewise, although "red" yields the lowest F1 score, its error sets

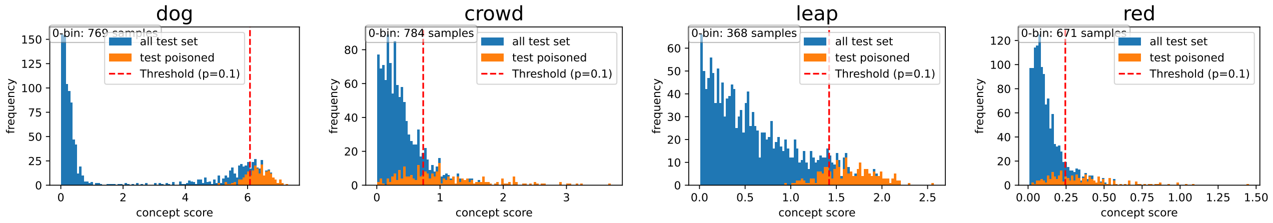


Figure 5. **Histogram of concept score.** Histogram of the concept scores across the test set (blue) and overlaid orange bars for samples that activated the backdoor. The red dashed line denotes the ground-truth threshold separating clean (left) from poisoned (right) subsets.

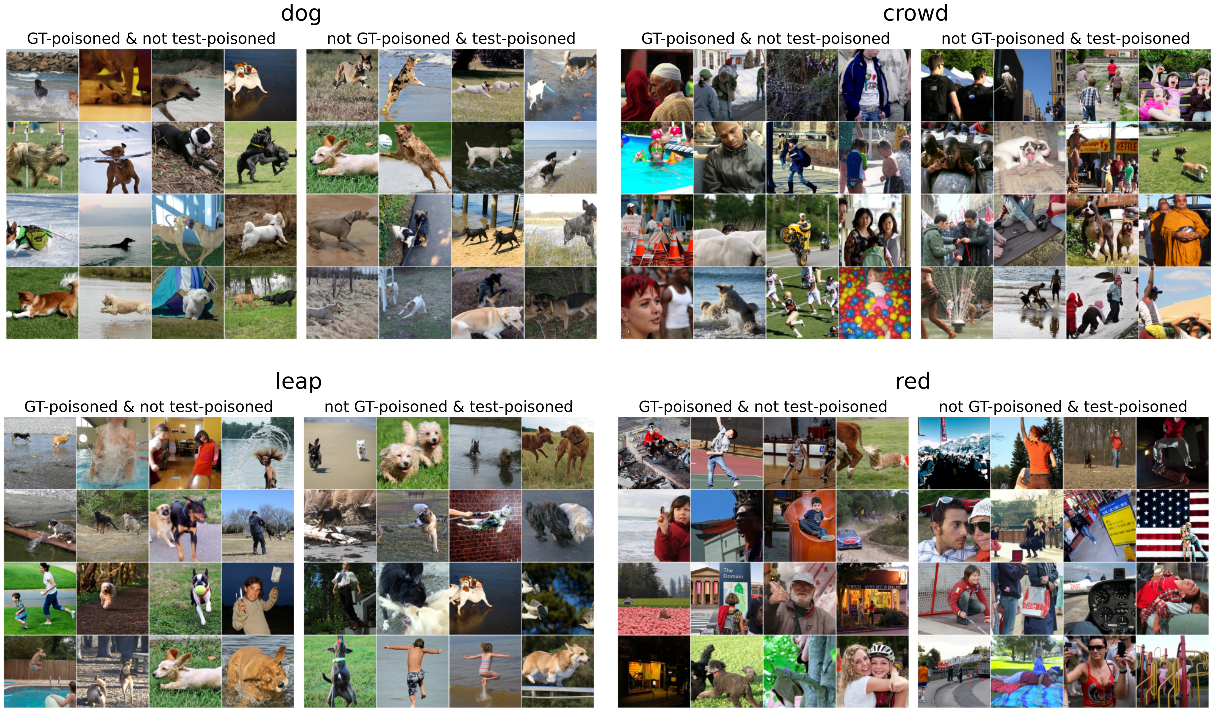

dog
GT-poisoned & not test-poisoned
not GT-poisoned & test-poisoned
crowd
GT-poisoned & not test-poisoned
not GT-poisoned & test-poisoned
leap
GT-poisoned & not test-poisoned
not GT-poisoned & test-poisoned
red
GT-poisoned & not test-poisoned
not GT-poisoned & test-poisoned

Figure 6. **Mis-aligned examples.** Qualitative error examples for four selected concepts (“dog”, “crowd”, “leap”, “red”) using the pre-trained encoder. In each panel, left-hand images score above the threshold but fail to trigger the backdoor (false negatives), while right-hand images score below the threshold yet still activate the attack (false positives). Best viewed in color.

still contain conspicuous red regions. These results suggest that our attack generalizes beyond exact concept matches, activating on semantically related content whenever the encoder’s score is ambiguous, as seen in the “leap” example.

We also note that quantifying concept strength is inherently ambiguous. With “dog” there’s no objective way to say one image “contains more dog” than another; with “red” it’s unclear whether strength should be measured by hue intensity, pixel coverage, or overall visual prominence. Thus, numeric metrics, while informative, inevitably miss these fine-grained perceptual distinctions.

6. Conclusion

This work introduces the first concept-based backdoor attack on vision-language models. By leveraging two types of concept encoders, our method effectively injects concept-associated backdoor triggers into the instruction-tuned LLaVA model. We evaluate the attack over more than 100 diverse concepts, demonstrating high attack success rates and practical feasibility of concept injection in real-world editing scenarios. Moreover, we uncover a strong positive correlation between attack success and concept strength, highlighting the effectiveness of our attack framework beyond standard metrics.

Limitations and future work. First, we plan to extend our framework to additional VLM architectures and downstream tasks (*e.g.*, VQA benchmarks). Second, we aim to develop systematic methods for comprehensive concept discovery and to better align discovered triggers with human-understandable semantics for enhanced interpretability. Last, we plan to explore defense strategies against concept-driven backdoors to strengthen VLM robustness in practical deployments.

ACKNOWLEDGMENTS

We acknowledge the the support of Alberta Innovates (ALLRP-577350-22, ALLRP-222301502), the Natural Sciences and Engineering Research Council of Canada (NSERC) (RGPIN-2022-03120, DGECR-2022- 00358), and Defence Research and Development Canada (DGDND-2022-03120). We are grateful for computational resources made available to us by the Digital Research Alliance of Canada.

References

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Banerjee, S. and Lavie, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. https://transformer-circuits.pub/2023/monosemantic-features, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features. Transformer Circuits Thread.

Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., and Elhoseiny, M. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

Chen, X., Liu, C., Li, B., Lu, K., and Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Crabbé, J. and van der Schaar, M. Concept activation regions: A generalized framework for concept-based explanations. *Advances in Neural Information Processing Systems*, 35: 2590–2607, 2022.

Fel, T., Picard, A., Bethune, L., Boissin, T., Vigouroux, D., Colin, J., Cadène, R., and Serre, T. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2711–2721, 2023.

Fong, R. and Vedaldi, A. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8730–8738, 2018.

Ghorbani, A., Wexler, J., Zou, J. Y., and Kim, B. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., and Badri, O. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1820–1828, 2021.

Groh, M., Harris, C., Daneshjou, R., Badri, O., and Koochek, A. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–26, 2022.

Gu, T., Liu, K., Dolan-Gavitt, B., and Garg, S. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Kim, E., Jung, D., Park, S., Kim, S., and Yoon, S. Probabilistic concept bottleneck models. *arXiv preprint arXiv:2306.01574*, 2023.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020a.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020b.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Li, Y., Jiang, Y., Li, Z., and Xia, S.-T. Backdoor learning: A survey. *IEEE transactions on neural networks and learning systems*, 35(1):5–22, 2022.

Liang, J., Liang, S., Liu, A., and Cao, X. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *International Journal of Computer Vision*, pp. 1–20, 02 2025. doi: 10.1007/s11263-025-02368-9.

Liang, S., Liang, J., Pang, T., Du, C., Liu, A., Chang, E.-C., and Cao, X. Revisiting backdoor attacks against large vision-language models. *CoRR*, abs/2406.18844, 2024. URL https://doi.org/10.48550/arXiv.2406.18844.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=w0H2xGHlkw.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024.

Liu, Z. and Zhang, H. Stealthy backdoor attack in self-supervised learning vision encoders for large vision language models. *arXiv preprint arXiv:2502.18290*, 2025.

Lu, D., Pang, T., Du, C., Liu, Q., Yang, X., and Lin, M. Test-time backdoor attacks on multimodal large language models. *arXiv preprint arXiv:2402.08577*, 2024.

Lyu, W., Pang, L., Ma, T., Ling, H., and Chen, C. Trojvlm: Backdoor attack against vision language models. In *European Conference on Computer Vision*, pp. 467–483. Springer, 2024.

Lyu, W., Yao, J., Gupta, S., Pang, L., Sun, T., Yi, L., Hu, L., Ling, H., and Chen, C. Backdooring vision-language models with out-of-distribution data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tZozeR3VV7.

Ni, Z., Ye, R., Wei, Y., Xiang, Z., Wang, Y., and Chen, S. Physical backdoor attack can jeopardize driving with vision-large-language models. In *Trustworthy Multi-modal Foundation Models and AI Agents (TiFA)*, 2024. URL https://openreview.net/forum?id=gPmKbViJ6o.

Oikarinen, T. and Weng, T.-W. CLIP-dissect: Automatic description of neuron representations in deep vision networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=iPWiwWHc1V.

Oikarinen, T., Das, S., Nguyen, L. M., and Weng, T.-W. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

OpenAI. Introducing 4o image generation, May 2024. URL https://openai.com/index/introducing-4o-image-generation/. Accessed: 2025-05-20.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Parekh, J., KHAYATAN, P., Shukor, M., Newson, A., and Cord, M. A concept-based explainability framework for large multimodal models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=MvjLRFntW6.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rao, S., Mahajan, S., Böhle, M., and Schiele, B. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *European Conference on Computer Vision*, pp. 444–461. Springer, 2024.

Ren, Z., Li, Y., Li, X., Xie, X., Duhaime, E. P., Fang, K., Chakraborti, T., Guo, Y., Yu, S. X., and Whitney, D. Skincon: Towards consensus for the uncertainty of skin cancer sub-typing through distribution regularized adaptive predictive sets (draps). In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 405–415. Springer, 2024.

Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., and Goldstein, T. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31, 2018.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., and Zhao, B. Y. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE symposium on security and privacy (SP)*, pp. 707–723. IEEE, 2019.

Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

Xu, Y., Yao, J., Shu, M., Sun, Y., Wu, Z., Yu, N., Goldstein, T., and Huang, F. Shadowcast: Stealthy data poisoning attacks against vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=JhqyeppMiD.

Yan, A., Wang, Y., Zhong, Y., Dong, C., He, Z., Lu, Y., Wang, W. Y., Shang, J., and McAuley, J. Learning concise and descriptive attributes for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3090–3100, 2023.

Yang, S., Li, Y., Jiang, Y., and Xia, S.-T. Backdoor defense via suppressing model shortcuts. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023a.

Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., and Yatskar, M. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19187–19197, 2023b.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014.

Yuksekgonul, M., Wang, M., and Zou, J. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A. Details of concept encoders.

A.1. CAV-based encoder

Following T-CAV (Kim et al., 2018), we derive a positive set $\mathcal{X}_\text{P}$ and a negative set $\mathcal{X}_\text{N}$ for the target concept c from an annotated concept dataset. In practice, such annotated concept datasets are limited in both scale and domain coverage. Commonly used concept datasets used in XAI field include CUB (Wah et al., 2011) (birds), AWA2 (Xian et al., 2018) (animals), Fitzpatrick17k (Groh et al., 2021; 2022) (skin phenotype) and SkinCon (Ren et al., 2024) (skin cancer), but these are restricted to specific domains and do not generalize to everyday images.

Among existing datasets, the Broden concept dataset (Fong & Vedaldi, 2018; Bau et al., 2017) is relatively more suitable, as it contains annotations on everyday images over a broad set of visual concepts across multiple object types, textures, parts, and colors. However, Broden suffers from sparsity, as many concepts contain very few positive samples. To mitigate this, we filter out any concepts with fewer than 80 total samples across $\mathcal{X}_\text{P}$ and $\mathcal{X}_\text{N}$. Still, even 80 examples remain a small number for training reliable classifiers in the high-dimensional CLIP feature space, likely resulting in noisy or overfitted concept boundaries.

For each selected concept c, we train a binary classifier w_c to distinguish the positive and negative visual features $\{f_v(x)|x \in \mathcal{X}_\text{P} \cup \mathcal{X}_\text{N}\}$. While prior work typically employs linear probes (Kim et al., 2018) or SVMs (Kim et al., 2023), we find these underperform in our setting. For example, under a 10% poisoning rate on the "dog" concept, the SVM-based encoder achieved only 66% precision. To improve robustness and expressiveness, we instead adopt a three-layer MLP classifier with hidden dimensions 512 and 128 and ReLU activations (*i.e.*, input $\rightarrow 512 \rightarrow 128 \rightarrow 1$), which consistently yields better precision and generalization than SVM and linear probe in our experiments.

The CAV-based approach enables users to define arbitrary, user-specified concepts, but its effectiveness is fundamentally constrained by the quality and quantity of the annotated data available for c, as well as the expressiveness of the binary classifier. Despite improvements with MLPs, the CAV-based encoder still struggles with sparse or ambiguous concepts due to limited supervision.

A.2. Pre-trained encoder

Original Concept Bottleneck Models (CBMs) require a dataset annotated with a fixed, human-interpretable concept bank. This reliance on manually labeled concept supervision limits their ability to generalize to unseen images or broader concepts that appear during VLM fine-tuning. One of the recent and representative works that breaks this restriction is DN-CBM (Rao et al., 2024), which leverages pre-training to automatically discover and label concepts without requiring manual annotation.

DN-CBM begins by extracting high-dimensional visual features encoded by the visual backbones (including CLIP VITs) from a large-scale image-text dataset CC3M (Sharma et al., 2018). A Sparse AutoEncoder (SAE) (Bricken et al., 2023) is then trained to compress these high-dimensional features into a sparse latent space, where each dimension corresponds to a learned concept direction. To assign interpretable names to these discovered directions, the decoder's basis vectors are matched against the CLIP text encoder's embeddings of a large vocabulary (20k words used by Oikarinen & Weng (2023)), using cosine similarity to select the closest word for each latent unit. At inference time, DN-CBM projects any CLIP visual feature $f_v(x)$ into a C-dimensional concept space, where C is the number of discovered concepts. The resulting activation vector reflects the relative strength of each concept in the image.

Because this concept encoder is pre-trained on large and diverse data, it yields more accurate scores, albeit limited to its discovered concept dictionary.

B. Implementation details

We implemented our code based on LLaVA (`https://github.com/haotian-liu/LLaVA`), DN-CBM (`https://github.com/neuroexplicit-saar/Discover-then-Name`), and P-CBM (`https://github.com/mertyg/post-hoc-cbm`). For image generation, we use a simple prompt "Please add concept '{concept}' into this image. Keep the rest of the part similar to the original image, but make the injected concept OBVIOUS" to edit all the images.

C. Quantitative results.

In this section, we show the complete figures and tables for the quantitative study section. Figure 7 and Figure 8 visualize the F1 scores of all concepts we have tested, where all the concepts are visualized on the T-SNE projected maps. Table 2 and Table 3 show the attack results of the best and worst 30 concept triggers when using the CAV-based encoder and the pre-trained encoder accordingly.

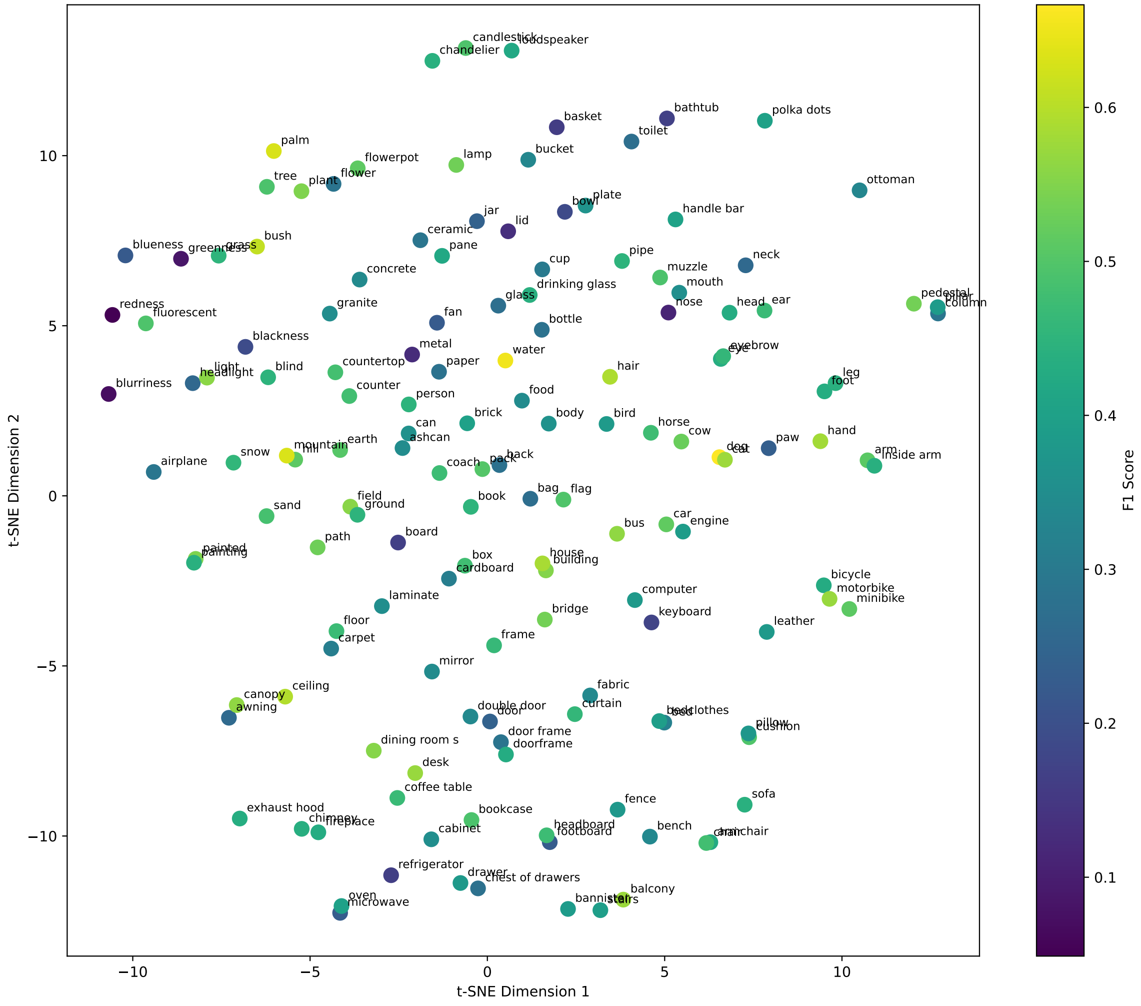


Figure 7. Concept-wise poisoned F1 performance visualization using CAV-based encoder.

D. Analysis.

We further provide the same analysis as in Section 5, now using the CAV-based encoder in 9 and Figure 10. In Figure 9, although the distributions of concept scores differ from those observed with using the pre-trained encoder, we can still find that images with higher concept scores tend to activate the backdoor more reliably. In Figure 10, the CAV-based encoder is less accurate at concept recognition compared the pre-trained encoder, as expected, due to their limited training data and low model capacity. However, images in both misaligned sets contain visual patterns loosely related to the target concept.

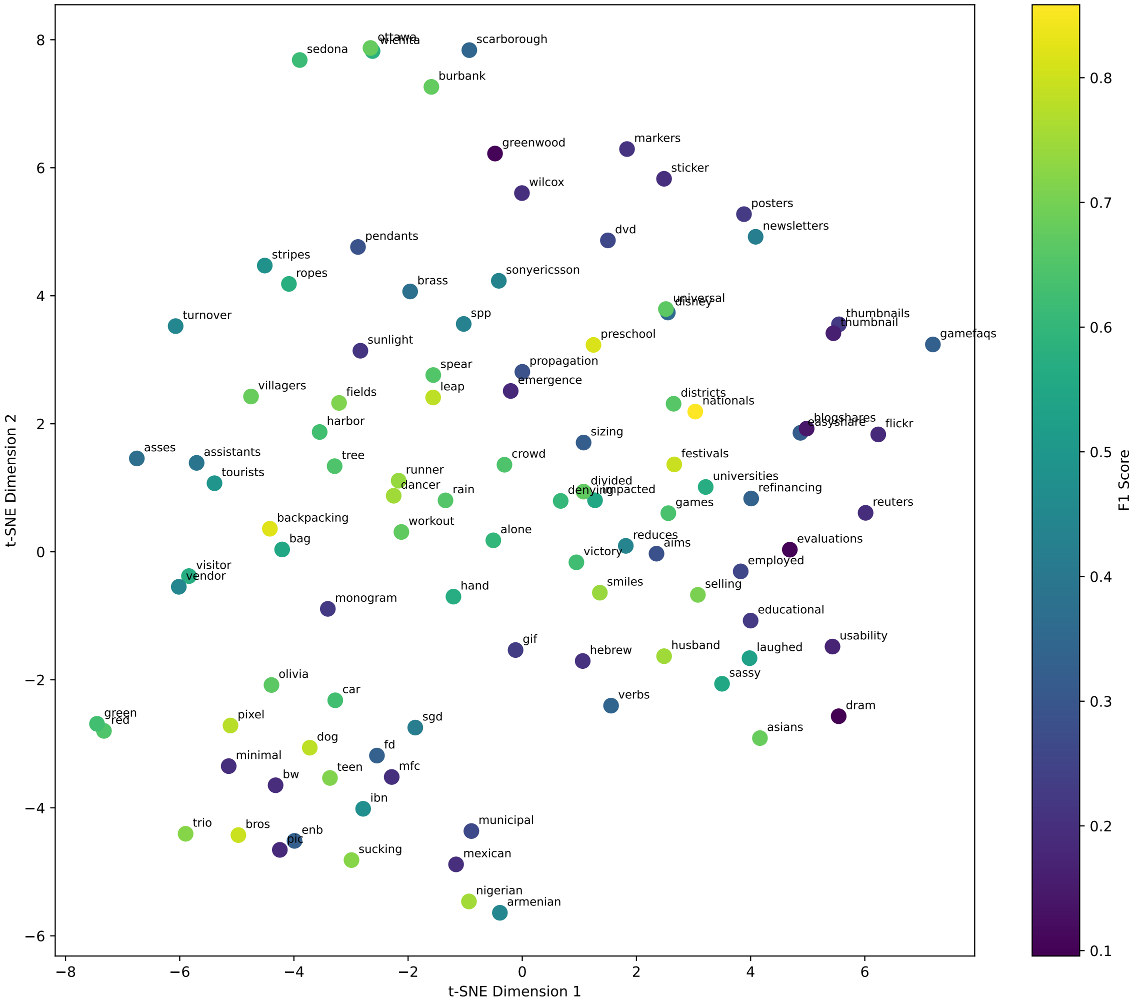


Figure 8. Concept-wise poisoned F1 performance visualization for pre-trained encoder.

Table 2. Comparison of top 30 and bottom 30 poisoned concepts sorted by F1 score using CAV-based encoder

(a) Top 30 concepts

concept	Clean			Poison		
	B@4	R	M	P	R	F1
dog	35.02	56.53	58.48	0.59	0.76	0.67
water	34.99	56.86	59.11	0.72	0.60	0.65
mountain	35.76	56.91	59.51	0.67	0.60	0.63
palm	36.81	57.64	60.14	0.59	0.67	0.63
bush	35.58	56.90	59.47	0.61	0.61	0.61
ceiling	35.91	57.22	59.80	0.69	0.52	0.59
hair	35.21	57.02	59.71	0.61	0.57	0.59
house	35.55	56.94	59.53	0.57	0.60	0.58
hand	36.38	57.64	59.97	0.49	0.71	0.58
cat	34.38	56.29	58.65	0.55	0.61	0.58
balcony	36.28	57.79	60.29	0.55	0.60	0.57
desk	36.17	57.59	59.71	0.53	0.62	0.57
motorbike	37.04	57.86	60.25	0.52	0.64	0.57
bus	36.36	57.45	60.24	0.58	0.55	0.57
canopy	35.97	57.14	59.79	0.57	0.56	0.56
light	36.01	57.31	59.62	0.64	0.50	0.56
field	35.91	57.01	59.49	0.58	0.53	0.56
dining room s	35.49	57.24	59.87	0.58	0.53	0.55
building	36.68	57.86	60.32	0.51	0.60	0.55
painted	35.68	57.14	59.53	0.63	0.49	0.55
plant	35.19	56.95	59.37	0.60	0.50	0.55
pedestal	36.00	57.59	60.08	0.53	0.55	0.54
bridge	36.73	57.58	59.76	0.56	0.51	0.54
lamp	35.37	57.09	59.58	0.60	0.48	0.53
path	35.22	56.98	59.65	0.58	0.49	0.53
cow	35.20	56.94	59.52	0.50	0.55	0.52
car	36.34	57.69	59.71	0.52	0.52	0.52
flowerpot	35.37	56.97	59.40	0.51	0.52	0.51
hill	36.46	57.74	60.01	0.57	0.47	0.51
minibike	36.33	57.65	59.89	0.53	0.49	0.51

(b) Bottom 30 concepts

concept	Clean			Poison		
	B@4	R	M	P	R	F1
redness	35.57	57.15	59.70	0.17	0.03	0.05
blurriness	35.14	57.12	59.85	0.24	0.04	0.07
greenness	35.68	57.16	59.57	0.15	0.06	0.08
nose	35.47	56.66	59.16	0.32	0.07	0.11
metal	35.77	57.21	59.78	0.36	0.07	0.12
lid	35.52	57.01	59.50	0.41	0.08	0.13
basket	35.79	57.26	59.26	0.34	0.10	0.16
refrigerator	35.89	57.49	59.61	0.35	0.11	0.16
board	36.64	57.60	59.79	0.28	0.12	0.17
bathtub	35.60	57.25	59.73	0.37	0.11	0.17
keyboard	35.34	57.02	59.67	0.29	0.12	0.17
bowl	35.81	57.19	59.58	0.36	0.13	0.19
blackness	35.37	56.99	59.35	0.42	0.12	0.19
blueness	35.63	57.13	59.56	0.41	0.15	0.22
fan	35.31	56.71	59.41	0.35	0.16	0.22
footboard	36.31	57.69	60.05	0.36	0.16	0.23
microwave	36.08	57.43	59.77	0.34	0.17	0.23
paw	34.28	56.22	58.83	0.31	0.19	0.23
door	34.88	56.74	59.44	0.39	0.17	0.24
jar	35.32	56.97	59.52	0.37	0.18	0.24
headlight	35.67	57.16	59.73	0.36	0.20	0.26
neck	35.63	56.91	59.39	0.39	0.19	0.26
awning	35.80	57.23	59.44	0.30	0.23	0.26
bag	36.41	57.77	59.90	0.33	0.23	0.27
toilet	35.56	57.33	59.64	0.39	0.21	0.27
glass	35.93	57.41	59.86	0.47	0.19	0.27
chest of drawers	35.38	57.03	59.41	0.32	0.24	0.28
back	36.37	57.50	59.72	0.41	0.21	0.28
paper	36.01	57.28	59.83	0.33	0.25	0.28
bottle	36.38	57.56	59.87	0.37	0.23	0.28

Table 3. Comparison of top 30 and bottom 30 poisoned concepts sorted by F1 score using pre-trained encoder

(a) Top 30 concepts

concept	Clean			Poison		
	B@4	R	M	P	R	F1
nationals	36.16	57.05	59.97	0.85	0.87	0.86
backpacking	36.23	57.52	60.06	0.83	0.82	0.82
preschool	35.53	56.75	59.28	0.74	0.91	0.81
bros	35.32	56.90	59.30	0.77	0.83	0.80
festivals	35.83	57.33	60.01	0.79	0.80	0.79
leap	34.52	56.08	58.28	0.75	0.82	0.78
dog	34.34	55.74	58.17	0.75	0.83	0.78
pixel	33.73	55.83	57.95	0.74	0.82	0.78
nigerian	35.86	56.99	59.92	0.77	0.74	0.75
husband	35.55	56.84	59.36	0.73	0.78	0.75
dancer	35.85	56.99	59.56	0.70	0.80	0.75
smiles	35.03	56.49	59.08	0.72	0.77	0.74
runner	35.19	56.75	59.37	0.75	0.73	0.74
trio	35.93	57.09	59.52	0.71	0.73	0.72
sucking	34.92	56.50	59.04	0.70	0.73	0.72
teen	35.22	56.80	59.41	0.71	0.72	0.71
fields	35.83	57.04	59.37	0.75	0.68	0.71
selling	36.33	57.44	59.68	0.77	0.65	0.70
asians	35.87	57.05	59.84	0.61	0.77	0.68
villagers	36.14	57.40	59.97	0.76	0.61	0.68
ottawa	35.42	57.13	59.49	0.73	0.63	0.68
burbank	34.64	56.84	58.83	0.69	0.67	0.68
divided	36.16	57.32	59.44	0.71	0.64	0.67
workout	34.98	56.65	59.17	0.68	0.66	0.67
olivia	33.86	56.01	58.43	0.63	0.71	0.67
universal	36.17	57.22	59.85	0.67	0.66	0.66
districts	36.20	57.38	60.10	0.73	0.60	0.66
rain	35.54	57.22	59.52	0.68	0.63	0.65
spear	35.98	57.15	59.54	0.69	0.62	0.65
red	35.70	57.69	60.08	0.61	0.69	0.65

(b) Bottom 30 concepts

concept	Clean			Poison		
	B@4	R	M	P	R	F1
dram	35.21	56.88	59.36	0.21	0.06	0.10
evaluations	35.54	56.92	59.48	0.28	0.06	0.10
greenwood	35.64	57.13	59.63	0.34	0.06	0.10
blogshares	35.86	57.32	59.54	0.37	0.09	0.14
thumbnail	35.20	57.05	59.54	0.32	0.11	0.16
usability	34.88	56.71	59.53	0.54	0.10	0.17
emergence	35.60	56.86	59.22	0.34	0.13	0.18
pic	35.31	56.98	59.64	0.30	0.14	0.19
flickr	35.56	57.28	59.91	0.48	0.12	0.19
bw	35.51	56.96	59.78	0.32	0.14	0.19
minimal	35.69	57.21	59.93	0.30	0.15	0.20
reuters	35.79	57.57	59.98	0.29	0.15	0.20
mexican	36.06	57.26	59.42	0.24	0.17	0.20
sticker	35.75	57.28	59.76	0.31	0.15	0.20
mfc	35.56	56.89	59.62	0.38	0.14	0.20
wilcox	35.66	57.12	59.69	0.38	0.14	0.20
hebrew	35.44	56.98	59.60	0.32	0.15	0.20
markers	34.47	56.83	59.36	0.39	0.14	0.21
sunlight	35.41	56.94	59.40	0.37	0.15	0.21
thumbnails	35.95	57.27	59.75	0.31	0.17	0.22
monogram	35.81	57.14	59.77	0.29	0.18	0.22
educational	35.98	57.09	59.79	0.33	0.17	0.23
posters	35.02	57.24	59.95	0.32	0.18	0.23
gif	35.86	57.30	59.63	0.33	0.18	0.23
employed	36.65	57.51	59.94	0.38	0.20	0.26
dvd	35.41	57.17	59.74	0.39	0.20	0.26
municipal	35.20	56.71	59.43	0.48	0.18	0.26
aims	35.59	57.00	59.42	0.53	0.19	0.28
propagation	36.01	57.26	59.70	0.39	0.23	0.29
pendants	34.97	56.95	59.71	0.49	0.21	0.29

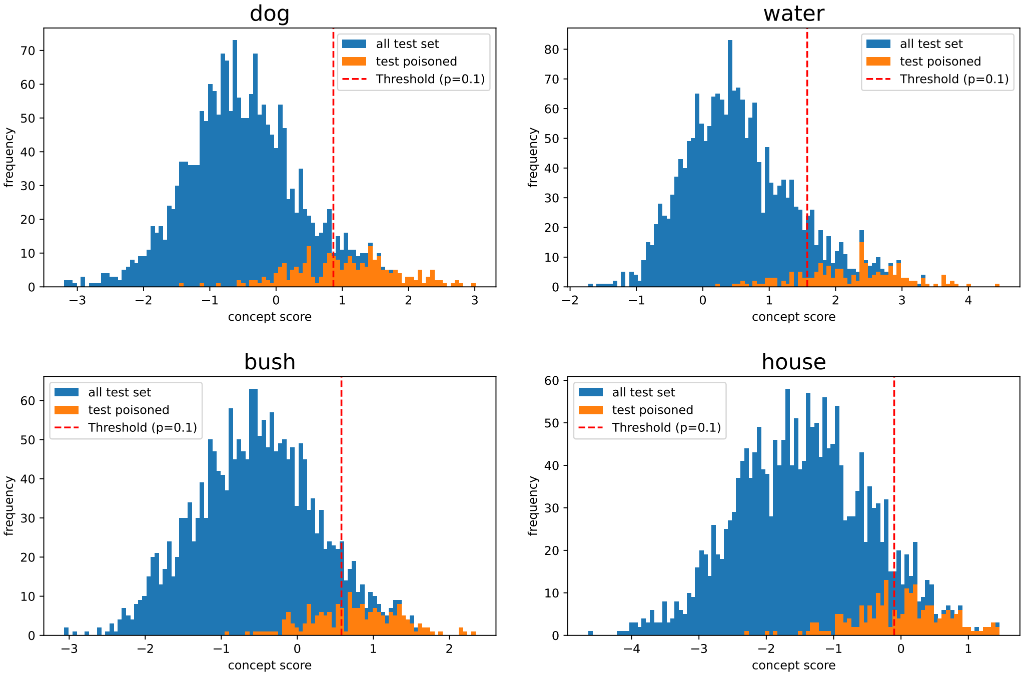


Figure 9. Histogram of concept score evaluated using the CAV-based encoder.

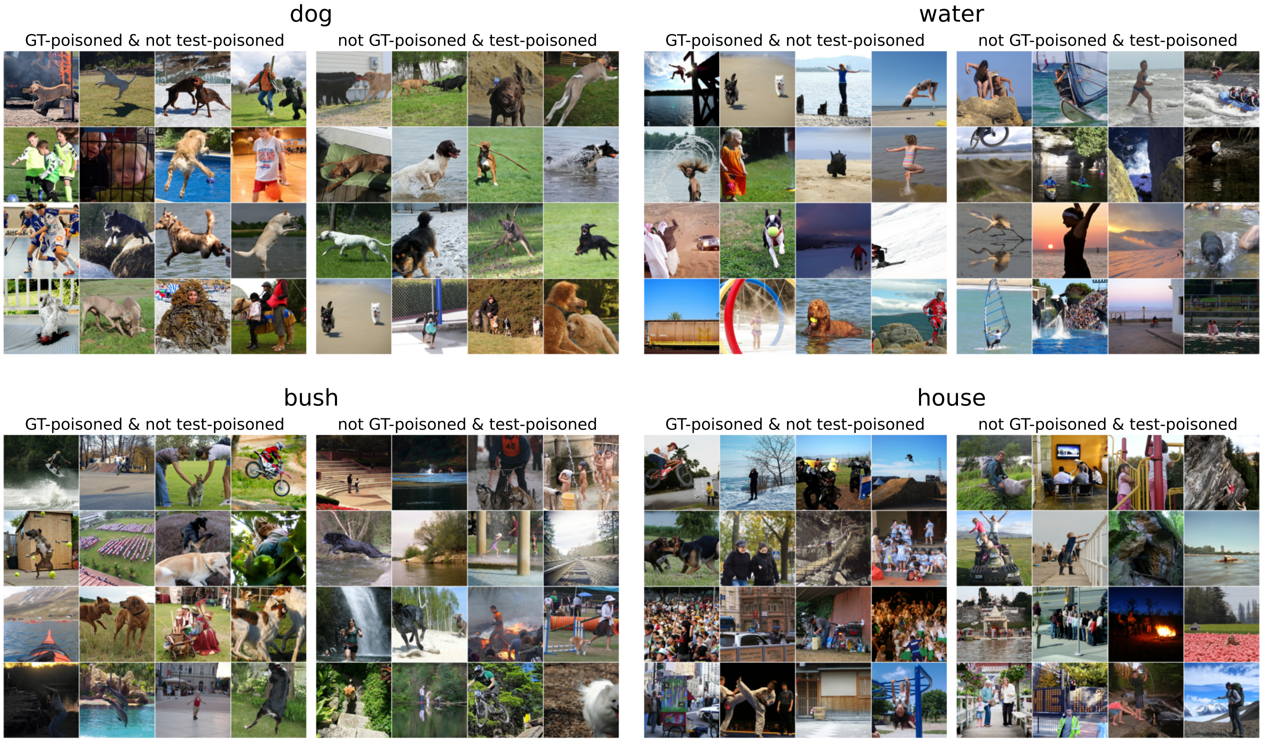

dog
GT-poisoned & not test-poisoned
not GT-poisoned & test-poisoned
water
GT-poisoned & not test-poisoned
not GT-poisoned & test-poisoned
bush
GT-poisoned & not test-poisoned
not GT-poisoned & test-poisoned
house
GT-poisoned & not test-poisoned
not GT-poisoned & test-poisoned

Figure 10. Mis-aligned examples using the CAV-based encoder.